

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s): Toru HISAMITSU, et al
Serial No.: 09/642,771
Filed: August 22, 2000
Title: WORD IMPORTANCE CALCULATION METHOD,
DOCUMENT RETRIEVING INTERFACE, WORD
DICTIONARY MAKING METHOD
Group: 2175

LETTER CLAIMING RIGHT OF PRIORITY

Honorable Commissioner of
Patents and Trademarks
Washington, D.C. 20231

August 22, 2000


Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, the applicant(s) hereby claim(s) the right of priority based on Japanese Patent Application No.(s) 11-237845 filed August 25, 1999.

A certified copy of said Japanese Application is attached.

Respectfully submitted,

ANTONELLI, TERRY, STOUT & KRAUS, LLP



Carl I. Brundidge
Registration No. 29,621

CIB/nac
Attachment
(703) 312-6600

日 本 国 特 許 庁

PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日

Date of Application:

1999年 8月25日

出 願 番 号

Application Number:

平成11年特許願第237845号

出 願 人

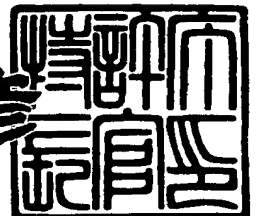
Applicant (s):

株式会社日立製作所

2000年 4月21日

特許庁長官
Commissioner,
Patent Office

近 藤 隆 彦



出証番号 出証特2000-3028608

【書類名】 特許願

【整理番号】 H99017431A

【提出日】 平成11年 8月25日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 17/30

【発明者】

 【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地
株式会社日立製作所中央研究所内

 【氏名】 久光 徹

【発明者】

 【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地
株式会社日立製作所中央研究所内

 【氏名】 丹羽 芳樹

【特許出願人】

 【識別番号】 000005108

 【氏名又は名称】 株式会社日立製作所

【代理人】

 【識別番号】 100075096

 【弁理士】

 【氏名又は名称】 作田 康夫

 【電話番号】 03-3212-1111

【手数料の表示】

 【予納台帳番号】 013088

 【納付金額】 21,000円

【提出物件の目録】

 【物件名】 明細書 1

 【物件名】 図面 1

 【物件名】 要約書 1

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 単語の重要度計算方法、文献検索インタフェイス、単語辞書作成方法

【特許請求の範囲】

【請求項 1】

文書集合に含まれる単語の重要度を計算する方法において、抽出すべき単語を含む部分文書集合と、全文書集合との乖離度を用いて該抽出すべき単語の重要度を計算することを特徴とする単語の重要度計算方法。

【請求項 2】

請求項 1 記載の単語の重要度計算方法において、上記乖離度は、上記部分集合と、上記全文書集合との距離 d と、上記部分集合と同程度の単語数を含み、かつ、上記全文書集合からランダム選出された部分文書集合と、上記全文書集合との距離 d' あるいは該 d' の推定値とを比較して求めることを特徴とする単語の重要度計算方法。

【請求項 3】

請求項 2 記載の単語の重要度計算方法において、二つの文書集合間の距離 d は、それぞれの文書集合における単語分布、すなわち全文書集合に含まれる各単語の出現確率を用いて計算することを特徴とする単語の重要度計算方法。

【請求項 4】

請求項 2 または請求項 3 のいずれかに記載の単語の重要度計算方法において、上記抽出すべき単語を含む部分文書集合に含まれる文書数が所定数より大きい場合、該部分文書集合からランダムサンプリングによりあらかじめ定めた数の文書を抽出し、該抽出された文書の集合と上記全文書集合との乖離度を用いて、該部分文書集合と該全文書集合の乖離度を推定することを特徴とする単語の重要度計算方法。

【請求項 5】

文書集合を特徴付ける単語を画面に提示する機能を持つ文献検索インタフェイスにおいて、全文書集合中に現れる各単語について、該単語を含む部分文書集合

中の単語分布と、該全文書集合中の単語分布との乖離度を用いて各々の単語の重要度を計算し、該重要度を画面に提示する単語の選択、配置、または、配色に反映させることを特徴とする文献検索インタフェイス。

【請求項 6】

文書集合を特徴付ける単語を画面に提示する機能を持つ文献検索インタフェイスにおいて、検索の結果得られる文書集合中に現れる各単語について、該単語を含む該検索の結果得られる文書集合の部分文書集合中の単語分布と、該検索の結果得られる文書集合中の単語分布との乖離度を用いて各々の単語の重要度を計算し、該重要度を画面に提示する単語の選択、配置、または、配色に反映させてることを特徴とする文献検索インタフェイス。

【請求項 7】

文書集合からあらかじめ与えた規則に従い重要単語を抽出する単語辞書作成方法において、全文書集合中に現れる各単語について、該単語を含む部分文書集合と、該全文書集合との乖離度を用いて各々の単語の重要度を計算し、該重要度に基づいて抽出すべき単語を選択することを特徴とする単語辞書作成方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、文書群中の単語または単語列の重要性を測る技術に係り、文献検索の支援、単語辞書の自動作成等に利用される。

【0002】

【従来の技術】

図 1 は、検索された文書内の「特徴単語」を提示するウィンドウを持つ文書検索システムの例であるが、右側のウィンドウには左側に示されている文書中の単語が選択されて表示されている。このような検索システムの例として、例えば、特開平10-74210「文献検索支援方法及び装置およびこれを用いた文献検索サービス」(文献 1)があげられる。

【0003】

また、影浦峽(他)、“Methodsof automatic term recognition: A review”、

Terminology、1998)(文献2)には、単語の重要度を計算する方法が記載されている。単語の重要度を計算する方法は、専門用語の自動抽出や、文献検索の際に文書の特徴付ける単語に重みをつけることを目的として、長い間研究されてきた。

【0004】

単語の重み付けに関しては、特定の文書内から重要語を抽出することを目的とするもの、全文書から重要語を抽出するのを目的とするものがある。前者についてもっとも有名なものは、tf-idfである。idfは、全文書数 N をある単語 w が現れる文書数 $N(w)$ で割ったものの対数、tfは単語の文書 d 内での出現頻度 $f(w, d)$ であり、tf-idfは、これらの積として

$f(w, d) \times \log_2(N/N(w))$ で表される。次のような、 $f(w, d)$ の平方根をとる等の変形がある

$f(w, d)^{0.5} \times \log_2(N/N(w))$ 他にもさまざまな変形があるが、tf-idfの基本的な性質として、「単語がより多く、より少ない文書に偏って出現するほど大きくなる」ように設定される。

【0005】

文献2には記述されていないが、この指標を特定の文書中での単語の重要度でなく、文書集合全体での単語の重要度を測る指標に拡張する自然な方法は、 $f(w, d)$ を、 w の全文書中での頻度 $f(w)$ に置き換えることである。

【0006】

全文書中から重要語を抽出するための方法の一つとして、単語の出現の偏りをより精密に捕えるために、注目する単語の、与えられた文書カテゴリごとの出現頻度の差異の偶然性を測り、偶然でない度合いが高いものを重要語としようという方法があり、尺度として χ^2 検定等が利用されているが、この場合、文書集合はあらかじめカテゴリに分類されている必要がある。

【0007】

これらとは別系統の研究として、自然言語処理の立場から重要語としてふさわしい語のまとまりを捕えようとする一連の研究がある。これらの研究においては、文法知識を用いて語の並びに制約を加えるとともに、隣り合う単語の共起の強

さをさまざまな尺度で測る方法が提案されている。尺度としては、(各点)相互情報量、対数尤度比等が利用されている。

【0008】

【発明が解決しようとする課題】

従来用いられてきた手法には、以下の問題があった：(1) tf-idf(もしくはその類似手法)の精度は不十分である。経験上語の頻度の寄与が大きすぎる傾向があり、例えば「する」のような一般的すぎる不用語の排除ができにくい。(2) 特定の語のカテゴリ間での分布の違いを比較する方法では、あらかじめ文書が分類されている必要があるが、この条件は一般に満たされない。(3) 隣り合う単語の共起の強さを利用する手法では、1単語のみの場合重要度が評価できない。 $n > 2$ について、拡張が自明ではない。(4) 従来は重要／非重要を分ける閾値の設定が困難かつ恣意的になりがちであった。本発明の目的は、このような問題の無い方法を提案することである。

【0009】

【課題を解決するための手段】

以下、タームとは、単語または単語列のことを示す。「タームの重要性」を、専門用語抽出や、情報検索の観点から言い替えると、あるタームが重要であるとは、そのタームがある程度のまとまった話題を想起させる、すなわち *informative* または *domain-specific* であることと言える。これは情報検索の領域では *representative* と呼ばれ、この意味でのタームの「重要性」は、*representativeness* とも呼ばれる。このようなタームは、文書集合の内容を俯瞰するときに役立つと考えられるため、情報検索やその支援システムにおいては重要である。

【0010】

*representativeness*を測る際、従来の方法は注目するターム自体の分布に着目していた。しかし、tf-idfのような手法は簡易ではあるが精度が不足し、 χ^2 等の統計量を用いる手法では、一つのタームは例外を除きせいぜい数十回しか現れないため、多くのタームについて統計的に意味のある値を得るのは困難であり、それが精度の低下につながっていた。

【0 0 1 1】

本発明は、特定のタームの分布でなく、注目するタームと一緒にあらわれる単語の分布に注目する。これは、「タームの重要度は、そのタームといっしょに現れる単語の分布の偏りと関係がある」という作業仮説に立つものであり、あるタームが「重要である」ということを、「そのタームと共に現われる単語の単語分布が特徴的である」と解釈する。

【0 0 1 2】

そこで、本発明では、上記課題を解決するため、文書集合に含まれる単語の重要度を計算する際に、抽出すべき単語を含む部分文書集合中の単語分布と、もとなる文書集合中の単語分布との乖離度を用いることを特徴とする。特に、上記乖離度は、上記部分集合と上記文書集合との距離 d と、上記部分集合と同程度の単語数を含み、かつ、上記文章集合からランダム選出された部分文章集合と上記文章集合との距離 d' とを比較して求めることを特徴とする。

【0 0 1 3】

【発明の実施の形態】

以下では、任意のタームの *representativeness* を求める手法と、その情報検索システムへの応用を示す。まず、上記「課題を解決するための手段」の欄で述べた考え方を数学的に言い替えることにより、タームの *representativeness* を測る指標を導入する。すなわち、任意のターム W (単語または単語列) について、 W を含む文書すべての集合における単語分布と、全文書の単語分布の距離に着目する。具体的には、 W : ターム (任意の個数の単語からなる)、 $D(W)$: W を含む文書すべての集合、 D_0 : 全文書の集合、 $PD(W)$: $D(W)$ における単語分布、 P_0 : D_0 における単語分布、とすると、 W の *representativeness* $Rep(W)$ を、2つの分布 $\{PD(W), P_0\}$ の距離 $Dist\{PD(W), P_0\}$ に基づいて定義する。

【0 0 1 4】

単語分布間の距離の計測の方法としては、主要なものだけでも、(1) 対数尤度比 (log-likelihood ratio)、(2) Kullback-Leibler divergence、(3) transition probability、(4) vector-space model (cosign 法) 等が考えられるが、例えば対数尤度比を用いて安定した結果が得られることを確認している。全単語を $\{w$

1、。。。、 w_n }、 k_i と K_i を、単語 w_i が $D(W)$ 、 D_0 に出現する頻度として、対数尤度比を用いた場合の $PD(W)$ と P_0 の距離を以下で定義する。

【0 0 1 5】

【数 1】

数 1

$$\sum_{i=1}^n k_i \log \frac{k_i}{\#D(W)} - \sum_{i=1}^n k_i \log \frac{K_i}{\#D_0}$$

【0 0 1 6】

図2は、日経新聞1996年版の記事を用い、そこにあらわれるいくつかの語 W に対し、各語 W について、 $D(W)$ の含む単語数 $\#D(W)$ を横軸に、 $\text{Dist} \{PD(W), P_0\}$ を縦軸にプロットしたものである。ここでは、距離として対数尤度比を用いている。図2から見られるとおり、 $\#D(W)$ が近いターム同士で比較すれば、たとえば「米国」は「する」、「オウム」は「結び付ける」より $\text{Dist} \{PD(W), P_0\}$ の値が高く直感と合致する。しかし、このままでは $\#D(W)$ が離れたターム(これは概ね、二つのタームの頻度が大きく異なることと等価である)同士の representativeness を適切に比較することができない。なぜならば、一般に $\text{Dist} \{PD(W), P_0\}$ は、 $\#D(W)$ が大きくなるにつれて増加するからである。実際、「オウム」は「する」と $\text{Dist} \{PD(W), P_0\}$ の値が同程度となる。そこで特定のタームから離れて $\text{Dist} \{P, P_0\}$ の振る舞いを調べるため、さまざまな数の文書をランダムサンプリングし、その結果得られたさまざまな大きさの文書集合 D に対して計算した ($\#D$ 、 $\text{Dist} \{PD, P_0\}$) を、図2に「×」を用いてプロットした。これらの点は、(0、0)に始まり ($\#D_0$ 、0) に終わる一つのなめらかな曲線により良く近似できると思われる。以下、この曲線をベースライン曲線と呼ぶことにする。

【0 0 1 7】

$D = \phi$ のときと、 $D = D_0$ のときに $\text{Dist} \{PD, P_0\}$ が0となるのは定義から明らかであるが、 $\#D = 0$ 付近の挙動は、比較的全文書数が少ないとき(2、000文書程度)から、新聞1年分(3000、000文書程度)まで、全文書集合が様々な大きさの

場合にかなり安定して近似できることが確認できた。

【0018】

そこで、上記のさまざまな大きさの全文書集合において、ベースライン曲線が指数関数を用いた近似関数を用いて安定して精度良く求められる区間($1000 \leq \#D < 20000$)上で近似関数 $B(\cdot)$ を求め、 $1000 \leq \#D(W) < 20000$ を満たす W のrepresentativenessを、 $\text{Dist}\{PD(W), P0\}$ に、 $B(\cdot)$ による正規化を施した値： $\text{Rep}(W) = \text{Dist}\{PD(W), P0\} / B(\#D(W))$ により定義する（ただし、ここでいう単語は、記号や助詞、格助詞などの情報検索の検索語として確実に不要とみなされたものはすでに除いたものを指す。これらを含めた場合でも同様の手法が実現できるが、その場合は上記の数字は若干異なってくる）。

【0019】

ここで、「する」のように著しく $\#D(W)$ が大きい場合でも、上記のベースライン関数の有効域を用いることを可能にすることと、計算量を低減することを意図して、 $20,000 < \#D(W)$ となるような W に対しては、 $D(W)$ として150文書程度をランダム抽出し、 $1000 \leq \#D(W) < 20000$ を満たすようにしてから $\text{Rep}(W)$ を計算する。

【0020】

一方、上記の区間で求めたベースライン曲線の近似関数は、 $\{x \mid 0 \leq x < 1000\}$ で、値を大きめに見積もる傾向があるため、 $\#D(W) \leq 1000$ となる W については、正規化の結果 $\text{Rep}(W)$ は低めに出る。しかし、1000単語はほぼ新聞の2、3記事に相当するが、出現文書数とその程度のタームは我々の目的からの重要度は低いため、そのまま適用した。もちろん、別のベースラインを計算しておいてもよい。

ランダムサンプリングした文書集合 D における $\text{Dist}\{PD, P0\} / B(\#D)$ は、さまざまなコーパスにおいて、安定して平均 Avr がほぼ $1(\pm 0.01)$ 、標準偏差 σ が0.05程度であった。また、最大値が $\text{Avr} + 4\sigma$ を越えることはなかったので、あるターム W の $\text{Rep}(W)$ の値が、「意味のある値である」と判断するための閾値として、 $\text{Avr} + 4\sigma = 1.20$ を設ける。

【0021】

上記指標 $\text{Rep}(\cdot)$ は、(1) 数学的な意味付けが明瞭であり、(2) 高頻度タームと

低頻度タームの比較が自然にできる。(3) 閾値の設定が自然にできる。(4) 任意の長さのタームに対して適用できる。等の好ましい性質を持つ。

【0022】

本発明で提案する指標 $\text{Rep}(\cdot)$ の有効性は、実験によっても確認されている。日経新聞1996年版の記事中、総頻度が3以上の単語から20,000語を無作為抽出し、そのうちの2,000個を、検索内容の概観に現われることが「好ましい a」「どちらでもよい」「好ましくない d」の3種類に人手で分類した。該20,000語を何らかの方法でソートしたときに、各クラスに分類された語の、先頭からN位までの累積出現頻度グラフを、ランダムソート、頻度、「従来の技術」において述べた、全文書を対象としたtf-idfの変形版、すなわち、 N を全文書数、 $N(w)$ を w が現れる文書数、 $f(w)$ を w の全文書中での頻度として、 $f(w) \times 0.5 \times \log_2(N/N(w))$ を用いた。

【0023】

図7は、分類が“a”となったものの累積頻度を、ランダム、頻度、tf-idf、新指標のそれぞれを用いた場合で比較したものである。グラフから明らかに、ランダム<頻度<tf-idf<新指標の順で「好ましい」と分類される語の優先順位を上げる力が強い。改善はあきらかに有為である。図8は、分類が“d”となったものの累積頻度の比較であり、新指標の選別能力の優位性がより際立っている。頻度とtf-idfはランダムな場合と変わらず、「不要語」特定能力の低さを現している。このため、本発明で提案する指標は、不要語の同定にとりわけ有効であり、「高頻度かつrepresentative-nessの低い語」を選ぶことによる、stop-word listの自動作成や、文献類似度計算における語の重み付けの精度改善等への応用が期待される。

【0024】

これまでに述べたrepresentativenessを計算するためのシステム構成例を図3に示す。以下図3、4を用いてrepresentativenessの計算について述べる。301は記憶装置であり、ハードディスク等を用いて文書データ、各種のプログラム等を格納する。また、プログラムの作業用領域としても利用される。以下、3011は、文書データ。以下の例では日本語を用いるが、言語にはよらない。

3 0 1 2 は、形態素解析プログラム、文書を構成する単語を同定する。日本語の場合は分かち書き+品詞付け、英語の場合は原型還元等の処理を行う。この手法については特定しない。両言語とも、商用・研究用をとわずさまざまなシステムが公開されている。3 0 1 3 は、単語・文書対応付けプログラム。形態素解析の結果から、どの単語がどの文書に何回あらわれているか、逆にどの文書にどのような単語が何回あらわれているかを調べる。基本的には単語と文書をそれぞれ行・列とする行列の要素を計数により埋める作業であり、この手法については特定しない。3 0 1 4 は、単語・文書対応データベース (DB)。上記で計算された単語・文書対応データを記録する DB。3 0 1 5 は、representativeness 計算プログラム。図 4 にその詳細を示す、タームの representativeness を計算するプログラム。3 0 1 6 は、計算されたタームの representativeness を記録する DB。3 0 1 7 は、複数のプログラム間で共通に参照するデータを記録する領域である。3 0 1 8 は、作業用の領域である。3 0 2 は、入力装置、3 0 3 は、通信装置、3 0 4 は、メインメモリ、3 0 5 は、CPU、3 0 6 は、ディスプレイ、キーボード等より構成される端末装置、である。

【0 0 2 5】

図 4 は、3 0 1 5 の representativeness 計算プログラムの詳細である。以下、これを用いて、特定のタームの representativeness を求める手法を説明する。4 0 1 1 は、背景単語分布計算モジュールである。このモジュールは、一度だけ用いられ、各単語の全文中での頻度を記録する。すなわち、(数 1) と同じく、全単語を $\{w_1, \dots, w_n\}$ とし、 K_i を単語 w_i が全文書 D_0 中に出現する頻度として、 (K_1, \dots, K_n) を記録する。4 0 1 2 は、与えられた文書データに対してベースライン関数を推定するモジュールである。このモジュールも、はじめに一度だけ用いられる。次の基本的な要素の組み合わせで実現できる。すなわち、

(1) 文書集合が与えられた時、含まれる単語数が 1000 語前後から 20, 000 語前後になるような文書集合をランダムに、しかも含まれる単語数が 1000 から 20, 000 語の間でできるだけさまざまな値をとるように選び、それぞれの場合に (数 1) を用いて 4 0 1 1 で求めた背景単語分布との距離を計算する。

【 0 0 2 6 】

(2) (1) で得られた点群と最小 2 乗法等を用いてベースライン関数 $B(\cdot)$ を算出する。 $B(\cdot)$ は、単語数から正の実数への関数である。これらに関する方法は特定しない。標準的な手法が存在する。

【 0 0 2 7 】

4 0 1 3 は、部分文書集合抽出モジュールである。(「ターム $W = w_{n1} \dots w_{nk}$ があたえられたとき、単語・文書対応 DB 3 0 1 4 により、各単語 $w_{ni} (1 \leq i \leq k)$ を含む文書集合 $D(w_{ni}) (1 \leq i \leq k)$ を求め、すべての $D(w_{ni}) (1 \leq i \leq k)$ の共通集合を取って $D(W)$ をもとめる。」 \rightarrow) ターム $W = w_{n1} \dots w_{nk}$ があたえられたとき、単語・文書対応 DB 3 0 1 4 により、各単語 $w_{ni} (1 \leq i \leq k)$ を含む文書集合 $D(w_{ni}) (1 \leq i \leq k)$ を求める。単語・文書対応 DB 3 0 1 4 に単語の文書内の位置情報まで記録されていると仮定すれば、ターム $W = w_{n1} \dots w_{nk}$ を含む文書集合は、すべての $D(w_{ni}) (1 \leq i \leq k)$ の共通集合の、ある部分集合としてもとめることができる。単語・文書対応 DB 3 0 1 4 に単語の文書内の位置情報まで記録されていない場合は、近似として、すべての $D(w_{ni}) (1 \leq i \leq k)$ の共通集合を取って $D(W)$ とする。4 0 1 4 は、共起単語ベクトル計算モジュールである。再び単語・文書対応 DB 3 0 1 4 により、 $D(W)$ 内の各単語の頻度を計数し、各単語 w_i が $D(W)$ に出現する頻度 k_i を求める。4 0 1 5 は、分布間距離計算モジュールである。(数 1) と、4 0 1 1、4 0 1 4 で求めた単語頻度を用いて、背景単語分布と、 W を含む全文書 $D(W)$ 内の単語分布との距離 $\text{Dist} \{PD(W), P_0\}$ を求める。4 0 1 6 は、分布間距離正規化モジュールである。 $D(W)$ 中の単語数を $\#D(W)$ として、4 0 1 2 でもとめた $B(\cdot)$ を用いて、 $B(\#D(W))$ をもとめ、 $\text{Rep}(W) = \text{Dist} \{PD(W), P_0\} / B(\#D(W))$ により、 W の representativeness を求める。4 0 1 7 は、ランダムサンプリングモジュールである。4 0 1 3 で、 $D(W)$ に含まれる文書数が多すぎる場合、あらかじめ定めた数 (共有データ記憶領域 3 0 1 7 中に記録) を越える場合、あらかじめ定めた数の文書を選ぶために用いられる。この例では、文書数をパラメータとしているが、望ましい単語数をパラメータとし、その適当な近傍に単語数が収まるように文書をランダムサンプリングするように設定することも可能である。

【 0 0 2 8 】

図 5 は、本発明を文献検索支援のための検索内容表示に応用する場合の構成例である。本図は、(文献1)の文献検索支援方法図 1 において示された構成図に沿って、ナビゲーションウィンドウにおける特徴語表示に本発明を適用する場合の検索装置の構成を示したものである。(文献1)の文献検索支援方法と異なるのは、特徴語表示手段ルーチン 5 4 4 において、5 4 4 5 の representativeness チェックルーチンが加わること、および、5 4 4 1 特徴語抽出ルーチン、5 4 4 2 共起関係解析ルーチン、5 4 4 3 グラフ配置ルーチン、5 4 4 4 グラフ表示ルーチンにおいて、5 4 4 1 特徴語抽出ルーチン、5 4 4 3 グラフ配置ルーチン、5 4 4 4 グラフ表示ルーチンにおいて、5 4 4 5 の representativeness チェックルーチンを参照することである。representativeness チェックルーチンは、全文書集合における各語の representativeness について、要求に応じてその値を返すルーチンである。各語の representativeness は、あらかじめ図 4 で示したプログラムにより計算しておくことが可能である。

【 0 0 2 9 】

ユーザがキーボード 5 1 1 より検索キーワードを入力すると、検索インタフェイス 5 2 1 には、そのキーワードを含む文書の見出し等が表示され、5 2 2 特徴語表示手段には、検索結果となる文書集合から選ばれた特徴語が表示される。5 4 4 1 の特徴語抽出ルーチンにおいて(文献1)の方法で単語をまず選出する。この中には、先にのべたごとく、「する」や「この」のような一般語が混入しているが、例えば、その単語が出現する文書集合中の単語数が、例えば 1 0、0 0 0 語を超えるような単語について、5 4 4 5 の representativeness チェックルーチンにより representativeness を調べ、その値があらかじめ定めた値(例えば 1. 2)より低い語を排除することにより、高頻度不要語の表示を抑制できる。さらに、5 4 4 3 グラフ配置ルーチン、5 4 4 4 グラフ表示ルーチンで 5 4 4 5 の representativeness チェックルーチンを参照することにより、(文献1)の方法で定める各頻度クラスの中で、「語の表示が重なる場合に representativeness が高い語ほど手前に表示する」、「representativeness の高い語ほど文字色を濃く表示する」等の操作を行うことは容易であるから、representativeness の高い語をよ

り目立たせるような手段で表示し、ユーザインタフェースを改善することが可能である。さらに上では、各語のrepresentativenessを、あらかじめ図4で示したプログラムにより計算しておくこととしたが、入力キーワードごとに得られる検索結果文書の集合に対して、これを改めて全文書集合と考え、図4に示したプログラムによって、検索結果文書に含まれる各語に対し、その場でrepresentativenessを計算することも可能である。5445のrepresentativenessチェックルーチンをそのように設計した場合、同じ語でもキーワードごとにrepresentativenessが異なってくるため、より適切に、状況を反映した特徴語を表示することが可能である。

【0030】

図6は、単語の自動抽出に本発明を用いる場合の構成例である。601は記憶装置であり、ハードディスク等を用いて文書データ、各種のプログラム等を格納する。また、プログラムの作業用領域としても利用される。以下、6011は、文書データ。以下の例では日本語を用いるが、言語にはよらない。6012は、形態素解析プログラムで、文書を構成する単語を同定する。日本語の場合は分かち書き+品詞付け、英語の場合は原型還元等の処理を行う。この手法については特定しない。両言語とも、商用・研究用をとわずさまざまなシステムが公開されている。6013は、単語・文書対応付けプログラム。形態素解析の結果から、どの単語がどの文書に何回あらわれているか、逆にどの文書にどのような単語が何回あらわれているかを調べる。基本的には単語と文書をそれぞれ行・列とする行列の要素を計数により埋める作業であり、この手法については特定しない。6014は、単語・文書対応データベースDB。上記で計算された単語・文書対応データを記録するDB。6015は、抽出単語格納DB。6017は、representativeness 計算プログラム（詳細は図4）6018は、計算されたタームのrepresentativenessを記録するDB。6019は、複数のプログラム間で共通に参照するデータを記録する領域である。601Aは、最終的な抽出の候補となる単語または単語列を選び出すプログラムである。内容については特定しないが、通常、例えば「与えられた文書形態素解析した結果から、助詞、助動詞、接辞を除いた単語集合」としてよい。601Bは、601Aの選び出した候補から、文法知

識を用いて用語として不適切な語の並びを排除するフィルタである。例えば格助詞や助動詞等が先頭や末尾にくるものなどを排除する。内容については特定しないが、例えば(文献2)で紹介された論文中にいくつか例がある。601Bの選出した候補は、601Cにより、特定の指標に基づき重要度を計算し、それがあらかじめ定めた値より低いものを排除したり、重要度に従ってソートして出力する。ここでは、もっとも頻繁に用いられる指標の名前にしたがって、tf_idfフィルタプログラムと呼ぶが、実際に用いる指標は、tf_idf以外の任意の指標であってよい。6016は、作業用の領域である。602は、入力装置、603は、通信装置、604は、メインメモリ、605は、CPU、606は、ディスプレイ、キーボード等より構成される端末装置、である。一般の単語抽出方法においては、6017、6018は用いられない。601Cの出力に対して、6017、6018により各候補のrepresentativenessを参照し、あらかじめ定めた値(例えば1.2)よりその値が小さいものを排除する。さらに変形として、601Cで6017、6018を用いて直接各候補のrepresentativenessを参照し、representativenessのみを用いて用語候補の選別を行うことも考えられる。

【0031】

図6に述べた構成の単語の自動抽出方法を用いて、人工知能関係の1870本の論文の要旨から用語抽出を行う実験を行ったところ、601A、601Bにより約18000個の用語候補が抽出された。601Cで、representativenessのみを用いた場合と、まずtf-idfを用いて用語候補をソートし、その出力に対してrepresentativenessを用いて非重要語の除去を行う二つの場合を行ったところ、最終的に抽出された用語候補は両者とも等しく約5000語であるが、後者の場合の方が、頻度順に近い順序で用語が抽出される傾向にあるため、人間に提示して最終判断を仰ぐ場合は、見なれた語が先頭に近く現れる後者の方が、有る意味で自然であるともいえる。

【0032】

【発明の効果】

本発明で提案する representativeness を用いる事により、文書集合中のタームについて、(1) 数学的な意味付けが明瞭であり、(2) 高頻度タームと低頻度ターム

ームの比較が自然にできる。(3) 閾値の設定が自然にできる。(4) 任意の長さのタームに対して適用できる。ようなタームの *representativeness* 計算方法、すなわち、単語または単語列の重要度を計算する方法が実現でき、単語情報検索インタフェース、単語抽出システム等の精度の向上に役立てることができる。

【図面の簡単な説明】

【図 1】

特徴単語提示ウィンドウを持つ情報検索支援システムの例。

【図 2】

二つの単語分布の間の距離を示す例。

【図 3】

提案する単語の重要度計算方法を実現するための装置構成。

【図 4】

representativeness 計算プログラムの構成。

【図 5】

文献検索支援のための検索内容表示に *representativeness* を応用する場合の構成例。

【図 6】

単語の自動抽出に *representativeness* を応用する場合の構成例。

【図 7】

提案する単語の重要度が、検索結果の要約にふさわしいと判断される単語の優先順位を高める力を他の指標と比較する実験結果のグラフ。

【図 8】

提案する単語の重要度が、検索結果の要約にふさわしくないと判断される単語の優先順位を下げる力を他の指標と比較する実験結果のグラフ。

【符号の説明】

3 0 1 : 記憶装置

3 0 1 1 : 文書データ 3

0 1 2 : 形態素解析プログラム

3 0 1 3 : 単語・文書対応付けプログラム

- 3 0 1 4 : 単語・文書対応データベース (DB)
- 3 0 1 5 : representativeness 計算プログラム
- 3 0 1 6 : representativeness DB
- 3 0 1 7 : 共有データ記録領域
- 3 0 1 8 : 作業用の領域
- 3 0 2 : 入力装置
- 3 0 3 : 通信装置
- 3 0 4 : メインメモリ、
- 3 0 5 : CPU
- 3 0 6 : 端末装置、
- 4 0 1 1 : 背景単語分布計算モジュール
- 4 0 1 2 : ベースライン関数推定モジュール
- 4 0 1 3 : 部分文書集合抽出モジュール
- 4 0 1 4 : 共起単語ベクトル計算モジュール
- 4 0 1 5 : 分布間距離計算モジュール
- 4 0 1 6 : 分布間距離正規化モジュール
- 4 0 1 7 : ランダムサンプリングモジュール
- 5 4 4 : 特徴単語表示手段作動ルーチン
- 5 4 4 1 : 特徴語抽出ルーチン
- 5 4 4 2 : 共起関係解析ルーチン
- 5 4 4 3 : グラフ配置ルーチン
- 5 4 4 4 : グラフ表示ルーチン
- 6 0 1 : 記憶装置
- 6 0 1 1 : 文書データ。
- 6 0 1 2 : 形態素解析プログラム
- 6 0 1 3 : 単語・文書対応付けプログラム
- 6 0 1 4 : 単語・文書対応データベース DB
- 6 0 1 5 : 抽出単語格納 DB。
- 6 0 1 6 : 作業用の領域

6 0 1 7 : representativeness 計算プログラム

6 0 1 8 : representativeness D B。

6 0 1 9 : 共通データ記録領域

6 0 1 A : 候補単語列抽出プログラム

6 0 1 B : 文法フィルタ

6 0 1 C : フィルタプログラム

6 0 2 : 入力装置、

6 0 3 : 通信装置、

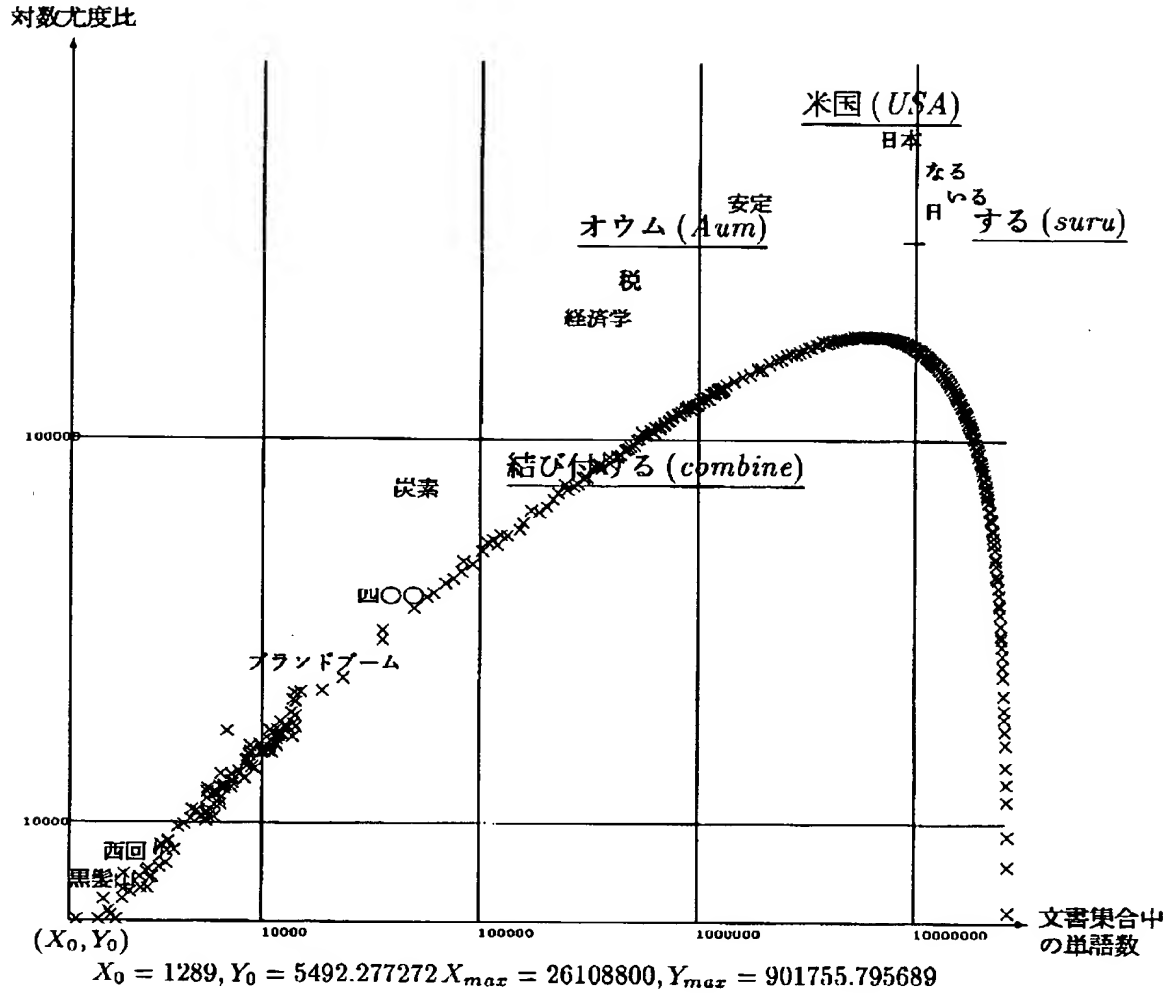
6 0 4 : メインメモリ、

6 0 5 : CPU

6 0 6 : ディスプレイ、キーボード等より構成される端末装置。

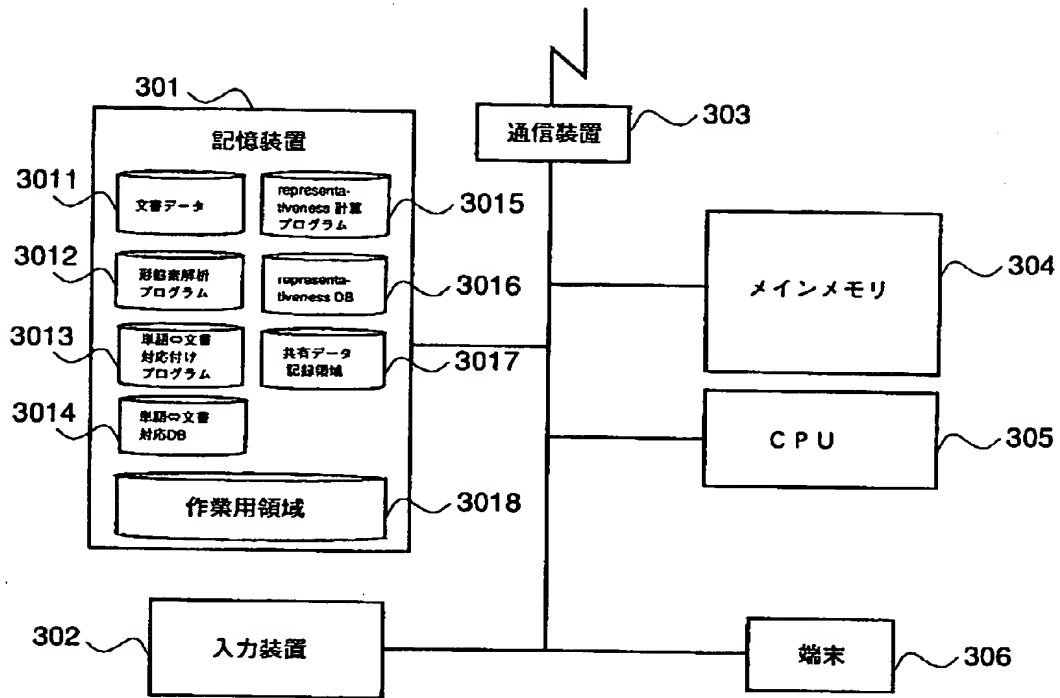
【図 2】

図 2



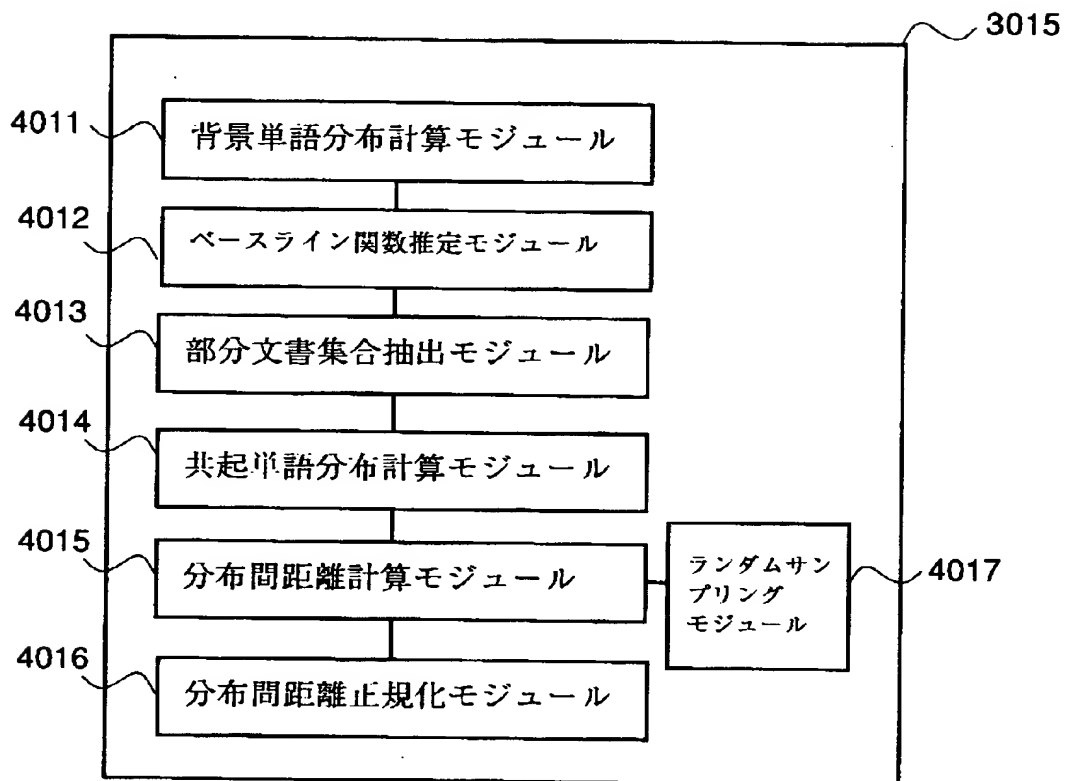
【図 3】

図 3

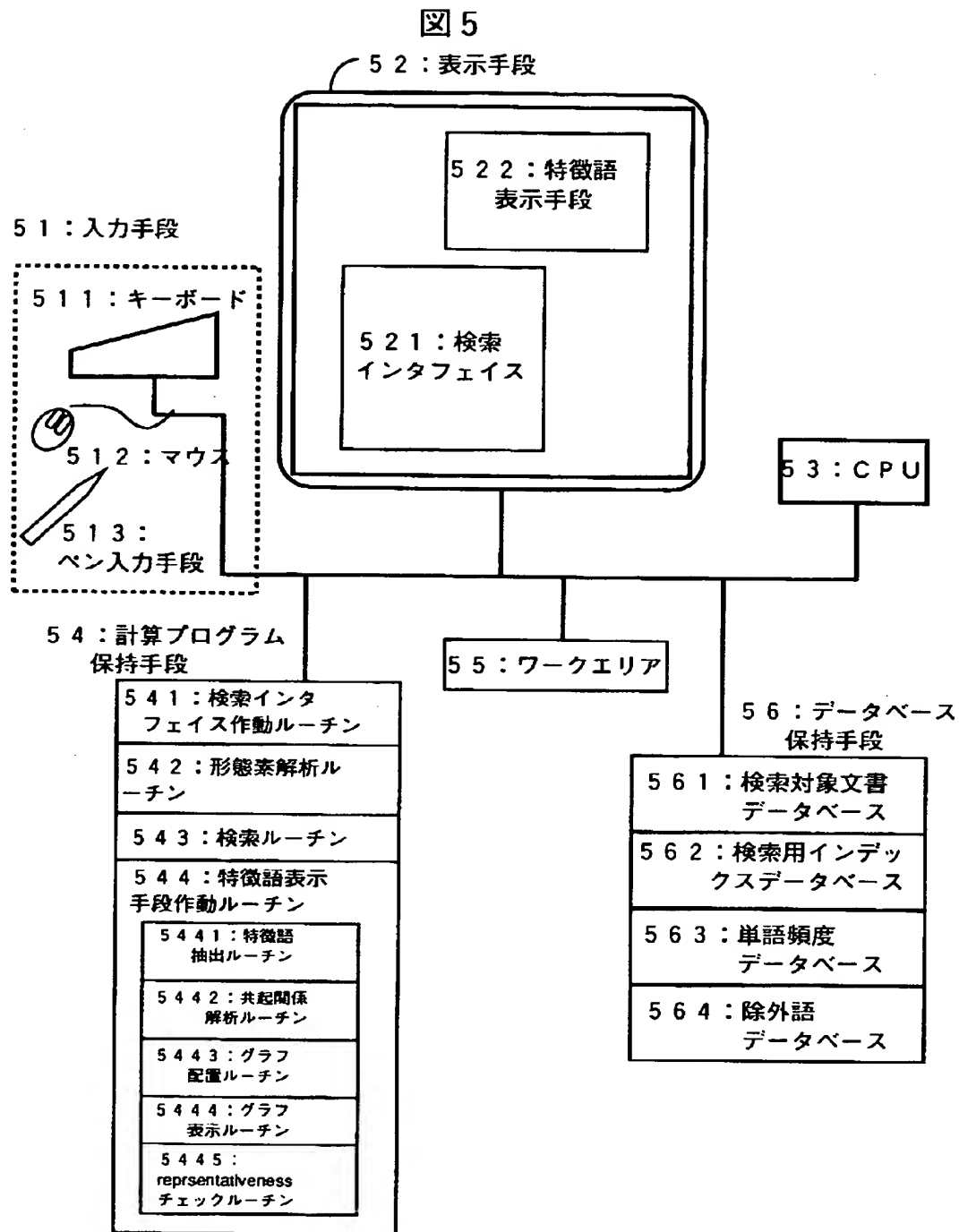


【図 4】

図 4
representativeness 計算プログラム

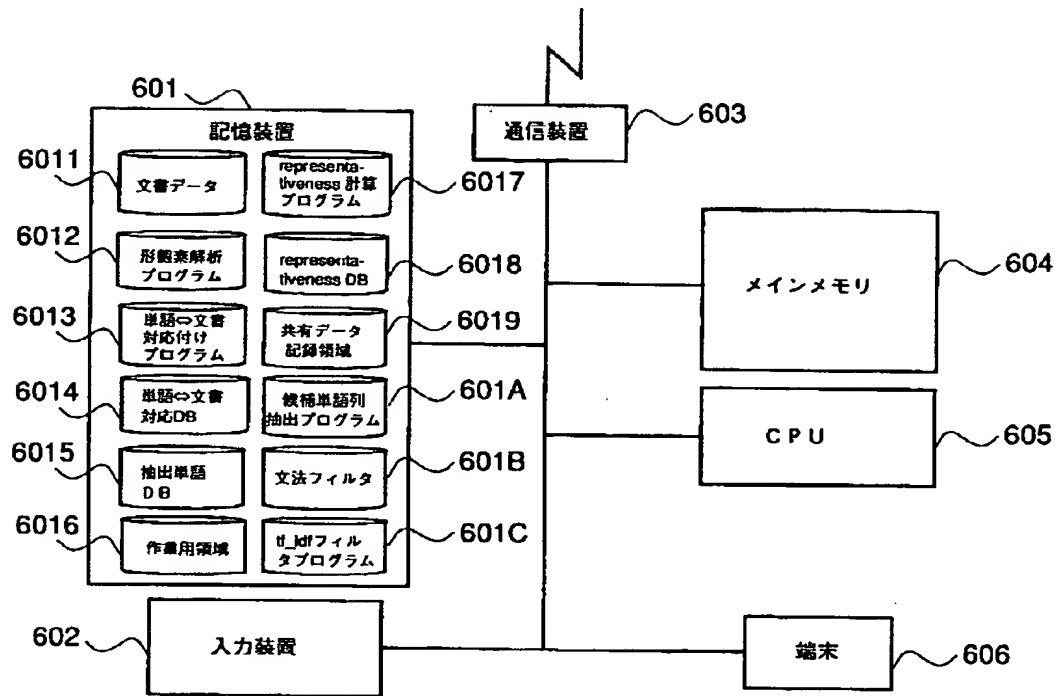


【図 5】



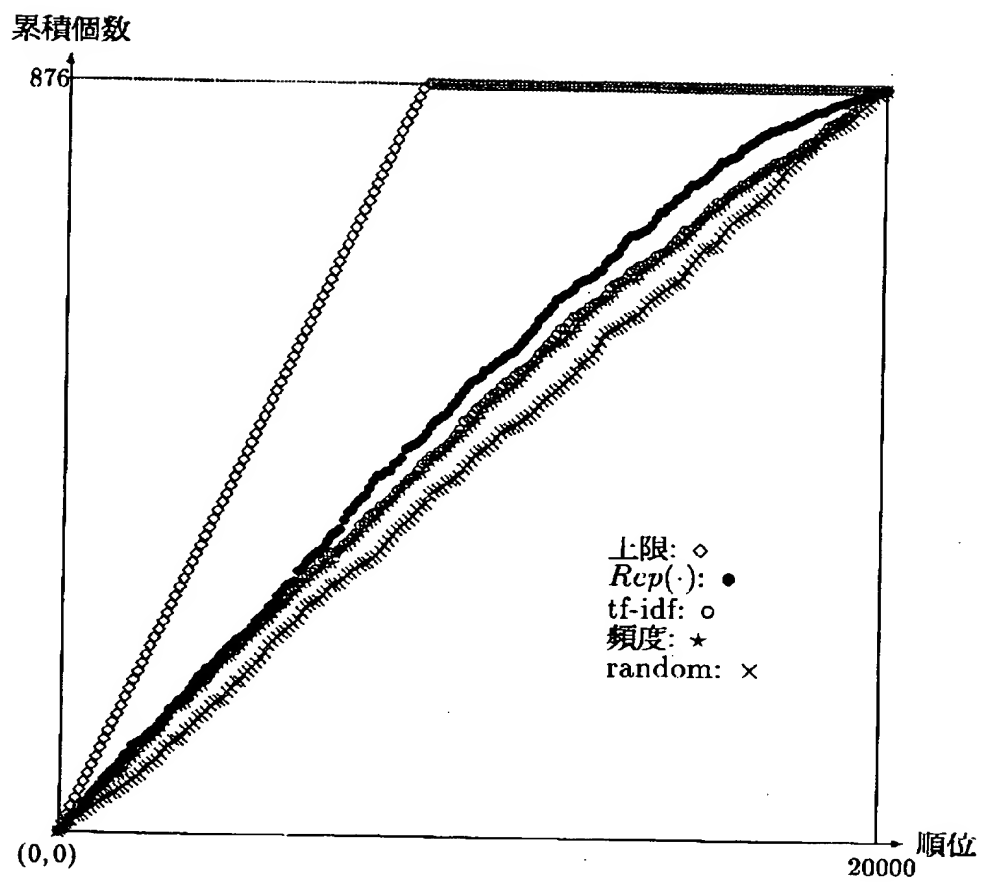
【図 6】

図 6



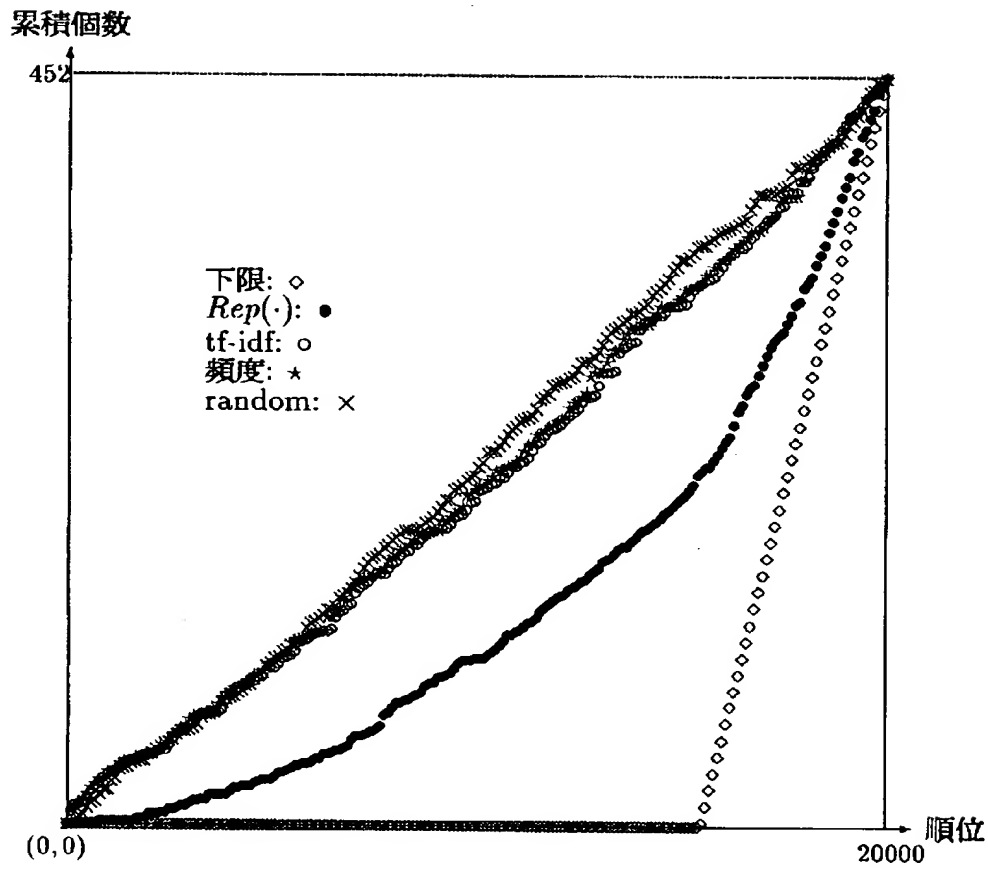
【図 7】

図 7



【図 8】

図 8



【書類名】 要約書

【要約】

【課題】 情報検索において重要である単語（列）を選択する方法においては、従来、高頻度一般語の除去ができない、重要／非重要を分ける閾値の設定が恣意的になりがちであるなど、いくつかの問題があった。本発明の目的は、このような問題の無い方法を提案し、情報検索支援の精度を向上させることである。

【解決手段】 上記の問題を、抽出すべき単語を含む部分文書集合中の単語分布と、もととなる文書集合中の単語分布との乖離度を、該部分文書集合中の単語数で正規化することにより解決する。

【選択図】 図 3

出 願 人 履 歴 情 報

識別番号 [000005108]

1. 変更年月日	1990年 8月31日
[変更理由]	新規登録
住 所	東京都千代田区神田駿河台4丁目6番地
氏 名	株式会社日立製作所